Selecting secondary measurements for soft sensor modeling using rough sets theory

Luo Jianxu
Institute of Automation
Shanghai Jiaotong University, 200030

Shao Huihe
Institute of Automation
Shanghai Jiaotong University, 200030

Abstract: **This paper contributed to introduce the rough sets theory for secondary measurements selection of soft sensor modeling. Using rough sets theory, a smaller set of secondary measurements containing the most information of the primary variable can be found. This method is used in Shanghai Refinery for the soft sensor estimating the propylene composition and got good result.**

1. INTRODUCTION

A major problem in product quality control is the lack of online quality sensors. Although in some cases analytical instruments are available, they possess substantial measurement delays, which make timely control impossible. Soft sensing technology is an economical and efficient solution for this problem [1]. Soft sensor is a modeling approach to estimating hard-to-measure process variables (primary variables) from easy-to-measure, online process sensors (secondary measurements).

A key stage of soft sensor modeling is the selection of secondary measurements, which have functional relationship with the estimated variable, from a large set of measurable candidate variables. The selection of secondary measurements includes the selection of types, numbers, and locations. Type selection usually depends on the mechanism analysis of the chemical process under investigation. The novel criterion is [2]: The selected secondary measurements should satisfy the requirements of sensitivity, accuracy, and robustness, etc.

The optimal secondary variable number is still an unresolved question. From the point of the view of cost considerations, mainly including investment and daily maintaining, a smaller set of secondary measurements is preferable. While from the viewpoint of covering more Information about the primary variable, people like to choose more measurements as secondary variables. Mejdell and Skogested [3] employed all available sensor signals as secondary variables to estimate the distillation column product composition, while Weng [4] chose only one temperature parameter as secondary variables for distillation column composition estimation.

Structured Singular Value analysis is a usually used approach to secondary measurements selection, which is adopted by Wisnewski[5], Luo[2] etc. Weng [4] used Karhunen-Loeve Transform for secondary measurements selection. These methods are usually implemented on simulation models, and are preferable for designing the optimal location of the secondary measurements.

This paper proposed a new approach to selecting the secondary measurements, by using rough sets theory (RST). RST was exposed by Pawlak as a method of set approximation in 1982 [6]. The integral part of RST is the construction of rule using reducts, which are particular subsets of the attributes providing classifications with the same quality of approximation as the full set of attributes[7].

The approach proposed in this paper makes use of the historical data collected from DCS, and the data is analyzed by the RST, then the subset of attributes, which can represent the primary variable, is found. This paper is organized as follows: the principles and concepts of the rough sets theory are introduced in part 2; the stages of the secondary measurements selection using rough set theory are represented in part 3; a practical application of the approach in Shanghai Refinery is introduced in part 4; at last the conclusion is showed in part 5.

2. ROUGH SET THEORY (RST)

2.1 Information system

An information system $S =< U, A, V, F >$ consists of U : a nonempty, finite set of objects (or cases), $U = \{x_1, \cdots x_n\}$; A : a nonempty, finite set of attributes, $A = \{A_1, \cdots A_n\}$; V : the domain set of A , $V = \{V_1, \cdots V_n\}$, where V_i is the domain of A_i ; $f : U \times A \to V$, an information function , $f(x_i, A_j) \in V_j$. An information system can be

0-7803-7268-9/01/$10.00 ©2001 IEEE.

represented with a decision table. In a decision table, the columns are the values of each attribute, and the rows are the objects.

Definition 2.1: For an information system $S =< U, A, V, f >$, let $B \subseteq A$, an *indiscernibility relation* is defined as follows:

$$IND(B) = \{(x_1, x_2) \in U \times U \mid b(x_1) = b(x_2), \forall b \in B\} = \bigcap_{b \in B} IND(b) \quad (2.1)$$

We say that x_1 and x_2 are indiscernible by set of attributes B if $b(x_1) = b(x_2)$ for every $b \in B$.

An order pair $AS =< U, IND(B) >$ is called an approximation space.

Definition 2.2: For any element $x \in U$, $B \subseteq A$, the *equivalence class* of x in relation B represented as $[x]_B$, and is defined as follows:

$$[x]_B = \{y \in U \mid (x, y) \in IND(B)\} \quad (2.2)$$

It is the equivalence class determined by x, which means each object in $[x]_B$ has the same attribute B value with that of x.

Definition 2.3: Let $X \subset U$, the *lower approximation* of X in AS for B is defined as :

$$\underline{B}(X) = \bigcup \{Y \in U / IND(B) \mid Y \subseteq X\} \quad (2.3)$$

and the *upper approximation* of X in AS for B is defined as:

$$\overline{B}(X) \bigcup \{Y \in U / IND(B) \mid Y \cap X \neq \Phi\} \quad (2.4)$$

For any $x \in \underline{B}(X)$, it is certain that it belongs to X with respect to B, and for any $x \in \overline{B}(X)$, we can only say that x is possible to belong to X.

Definition 2.4: Further more, we define the *positive region* and the *negative region* of X with respect to B:

$$POS_B(X) = \underline{B}(X), \quad NEG_B(X) = U - \overline{B}(X) \quad (2.5)$$

The positive region $POS_B(X)$ includes all objects in U which can be classified into classes of $IND(B)$ without error just based on the classification information in $IND(B)$.

2.2 Core and Reducts of attributes

Definition 2.5: Let $R \subseteq A$ be an equivalence relation, $r \in R$, the attribute r is *superfluous* if $IND(R) = IND(R - \{r\})$, otherwise r is *indispensable* in R.

If an attribute is superfluous, it can be removed from the information system, while an indispensable attribute carries the essential information about objects of the information system, and it should be kept to retain the characteristic of the information system.

R is *independent* if for every $r \in R$, r is indispensable.

Definition 2.6: For a subset of attributes $P \subseteq R$, if there exists $Q = P - \{r\}$, $Q \subseteq P$, satisfying $IND(Q) = IND(P)$, and Q is independent, then Q is called a *reduct* of P, denoted as $RED(P)$.

Definition 2.7 : The set of all indispensable attributes in P contained by all reduts is called the *core* of P, denoted as:

$$CORE(P) = \bigcap RED(P) \quad (2.6)$$

3. USING RST for SECONDARY VARIABLE SELECTING

3.1 Get initial information system using the sample data set

Consider MISO system. Suppose there are m pieces of sample data, in the form $\{x_1, x_2, x_n, \ldots y\}$. The sample data set is used to generate the information system: $S =< U, A, V, f >$, where U is finite universe of objects, here is the m pieces of data, $A = C \bigcup D$ is the set of finite attributes , $C = \{x_i\}$ is the set of condition attributes,(here is the input variable x_i), $D = \{y\}$ is the set of decision attributes,(here is the output variable y), $V = \bigcup_{r \in R} V_r$ is the set of all possible attribute values , V_r is the range of attribute r, $r \in R$, and $f : U \times R \to V$ is the information function. The initial information system can be represented with table 3.1.

Table1 3.1 initial information system

U	x_1	x_2	x_3	x_i	x_n	y
1	x_{11}	x_{12}	x_{13}	x_{1i}	x_{1n}	y_1
2	x_{21}	x_{22}	x_{23}	x_{2i}	x_{2n}	y_2
3	x_{31}	x_{32}	x_{33}	x_{3i}	x_{3n}	y_3
k	x_{k1}	x_{k2}	x_{k3}	x_{ki}	x_{kn}	y_k
m	x_{m1}	x_{m2}	x_{m3}	x_{mi}	x_{mn}	y_m

3.2 Discretization of continuous attribute values

There exist a great number of successful discretization algorithms and applications [8], such as equal-width-intervals, equal-frequency-intervals, Minimal Entropy Method and fuzzy method etc. In this paper, we apply fuzzy method to condition attributes, and equal-width-interval method to the decision attribute.

a. Fuzzy discretization of condition attribute value

For each condition attribute x_i, $(x_i \in \lfloor x_i^{\min}, x_i^{\max} \rfloor)$, $\lfloor x_i^{\min}, x_i^{\max} \rfloor$ is divided into m_i parts. Each partition point corresponds a fuzzy set $A_{ij}((j = 1,2,.....m_i)$, and the membership function value of the point is 1. The membership function of A_{ij} is Gaussian function.

Suppose x_{ij} is the value of the attribute x_i, let

$$\mu_{A_{is}}(x_{ij}) = \underset{k \in \{1,....m_i\}}{MAX}\{\mu_{A_{ik}}(x_{ij})\} \qquad (3.1)$$

Where A_{ik} is the fuzzy subset of x_i, and it is also corresponding to a discrete value k。Here, A_{is} has the maximal membership function value, then select s as the discrete value of x_{ij}.

b. Equal-width-interval discretization of decision attribute value

For decision attribute, $y \in \lfloor y_{\min}, y_{\max} \rfloor$, divide $\lfloor y_{\min}, y_{\max} \rfloor$ into n_i parts, each partition point is Y_j, $j \in (1,2,3..n_i)$。If y is a decision attribute value, seek Y_k, satisfied the following equation:

$$|y - Y_k| = \min\{|y - Y|\}, k \in (1,2,...n_i) \qquad (3.2)$$

k is the discrete value of y.

Substitute the continuous attribute values with discrete attribute values in table 3.1, and then we get the discrete information table. We define a confidence for each object as follows:

$$\mu_k = \prod_{i=1,...n} \mu_{A_{is}}^{(k)}(x_{ki}) \qquad (3.3)$$

Here, k is the kth object.

3.3 Choose secondary measurements set by removing the superfluous condition attributes

If an attribute is superfluous, it can be removed from the information system, while an indispensable attribute carries the essential information about objects of the information system, and it should be kept to retain the characteristic of the information system.

For attribute set c, $c = \{x_1, x_2, ...x_n\}$, if $\cap(c - \{x_i\}) = \cap c$, then x_i in c is superfluous. How to judge weather x_i is superfluous or not? If no incompatible objects (which means they have same condition attributes but different decision attribute) will appear after x_i is removed, then x_i is superfluous. Remove all superfluous attributes from attribute set c, and the left attributes are considered as secondary measurements.

4. APPLICATION

A schematic of a distillation column operated by Shanghai refinery is shown in Fig. 4.1.The whole system contains tower A, tower B, and tower C. The main product is propylene, whose composition in the top product and in bottom product of the tower C needs to be estimated. In the following contents, we'll concentrate on the estimation of the propylene in the top product, and apply the RST approach to selecting the secondary measurements for soft sensor predicting the propylene composition. All measurable candidate measurements are listed in the Table 4.1.

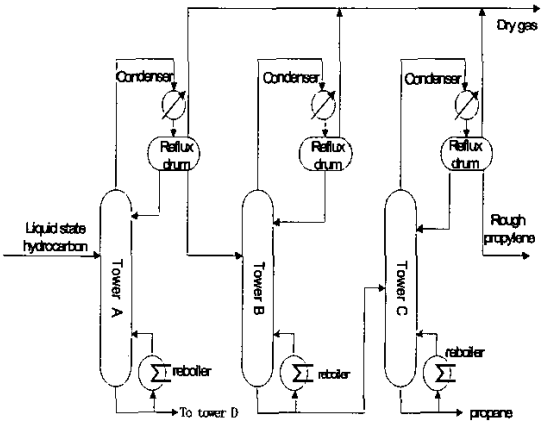

Fig 4.1 The schematic of the distillation column

TABLE 4.1 MEASUREBLE CANDIDATE MEASUREMENTS

Item	Label	Item	Label
Bottom pressure	PC04	Feed temperature	TD17
Feed flow rate of tower A	FC01	Top temperature	TD20
Feed flow rate	FC12	Reflux temperature	TD21
Reflux flow rate	FC15	Bottom temperature	TD22
Propylene flow rate	FC16	Reboiler temperature	TD23
Propane flow rate	FC18		

The propylene composition (denoted as AN1101) is sampled once every four hours and analyzed in laboratory. We collected 50 samples of the primary variable (the propylene composition) from laboratory analysis and the corresponding 50 samples of each measurement from DCS. These data are used for RST-based secondary measurements selection

The initial information system is established using the data samples as table 4.2 in appendix represents. According to the discretization method described in the part 3, we discretize the input variables (the candidate secondary measurements) into four attribute values, and the primary variable into three attribute values. Then the discrete information table is obtained as table 4.3, in appendix, shows.

For further processing of the decision table, we first combine the objects with the same condition attributes values and the same decision attribute values. Then for incompatible objects, keep the object with bigger confidence, remove the object with smaller confidence. In this case, after the processing, a table (table 4.4) of 49 objects is obtained.

Then judge which condition attribute is dispensable. The judge method is: remove the attribute from the decision table (a whole column of the decision table), if incompatible objects appear, then the attribute is indispensable, otherwise the attribute is dispensable. For example, judge whether FC01 is dispensable or not. Remove FC01 out of the decision table, there are not incompatible objects appearing in the new decision table, so FC01 is superfluous and can be removed. On the contrary, if FC02 is deleted from the decision table, (table 4.5 is the corresponding decision table) then object 13 and object 14 have same condition attributes but different decision attributes. They are incompatible. So FC02 is indispensable.

After all condition attributes are processed, the subset of indispensable condition attributes includes six attributes: FC12, TD20, TD21, FC15, TD23, and FC16. These variables are selected as secondary measurements. From the viewpoint of the mechanism analysis, the tower top temperature, the reflux temperature, the feed flow rate, and the top flow rate etc are most relevant to the top composition.

5. CONCLUSION

In this paper, we developed a new approach to selecting secondary measurements using rough sets theory. This approach can find a smaller set of secondary measurements from a large set of all available online measurements, and contain the most information about the primary variable. The application of this approach in Shanghai Refinery shows that it is a helpful method for secondary measurements selection in soft sensor modeling.

ACKNOWLEDGEMENTS

This paper is financially supported by the Chinese national ninth-five-year science and technology project foundation. The author also wishes to acknowledge Jiandong Xu, the senior engineer in Shanghai refinery, for his provision of live data.

REFERENCES

[1] Macvoy T J, "Contemplative stance for chemical process", *Automatica*, 1992, 28(2):441～442

[2] R.F.Luo and H.H. Shao, "Research of secondary measurements selection in inferential control", *Chinese control and decision annual conferenc*, 1992,6: 89～94

[3] Mejdell Thor and Sigurd Skogestad, "Composition estimator in a pilot-plant distillation column using multi-temperatures", *Ind. Eng. Res.* 1991, Vol.30: 2555～2564.

[4] Y.F.Weng, " Simulation, optimization and control of a distillation process", *Shanghai Jiaotong University doctoral dissertation.* Shanghai, 1997.

[5] P.A. Wisnewski, F.J. Doyle, C.J. Primus, "Measurement selection issues for the model predictive control of a Kamyr digester", Control Systems '96, Halifax, Canada, 1996, 153～156.

[6] Pawlak Z, "Rough Sets", International Journal of Information and Computer Science, 1982, 11(5):341～356

[7] Pawlak Z, "Rough Sets-Theoretical Aspects of Reasoning about Data", Dordrecht: Kluwer Academic Publishers, 1991, 68～162

[8] Robert Susmaga, "Analyzing Discretizations of Continuous Attributes Given a Monotonic Discrimination Function", *Intelligent Data Analysis*, 1997, 1:157～179

APPENDIX

TABLE 4.2 INITIAL INFORMATION SYSTEM

Sample	FC01 t/h	FC12 t/h	PC04 MPa	TD20 °c	TD21 °c	TD22 °c	TD17 °c	FC15 t/h	TD23 °c	FC16 t/h	FC18 t/h	AN01 V%
1	24.9	8.28	1.91	48.5	37.3	57.9	60.0	89.0	57.8	7.0	1.8	97.53
2	25.2	9.4	1.91	48.5	37.0	57.8	59.6	90.2	57.9	6.9	1.8	97.10
3	24.8	7.0	1.90	48.4	36.9	57.7	59.8	89.8	57.7	7.1	1.8	97.24
4	24.9	9.3	1.89	48.5	38.6	57.9	60.1	90.8	57.9	7.1	1.8	97.07
5	25.1	9.3	1.91	48.3	38.7	57.7	59.8	89.6	57.9	7.4	1.8	97.74
6	25.0	9.1	1.90	48.4	38.8	57.7	60.0	89.8	57.9	7.2	1.8	97.41
7	25.2	8.5	1.90	48.3	39.3	57.8	60.0	90.8	57.8	6.7	1.9	97.13
8	24.8	8.7	1.91	48.5	39.9	57.9	59.9	89.8	57.9	7.4	1.91	96.85
9	24.9	8.3	1.90	48.3	40.1	57.9	59.5	90.5	58.0	7.0	1.9	96.67
10	25.0	8.0	1.91	48.4	40.0	57.9	59.9	90.3	57.9	5.4	2.1	96.92
11	24.9	8.7	1.91	48.4	40.2	57.9	59.9	89.9	58.0	5.0	1.9	97.16
12	25.3	7.8	1.90	48.5	39.8	58.0	59.6	91.3	58.1	6.0	1.9	97.18
13	24.8	8.9	1.9	48.3	37.4	57.4	59.7	89.6	57.5	7.0	1.8	97.05
14	25.0	8.3	1.9	48.3	37.3	57.4	59.7	89.9	57.5	7.1	1.8	97.59
15	25.0	8.6	1.9	48.4	37.4	57.5	59.9	90.4	57.7	7.7	1.8	97.33
50	24.8	9.1	1.91	48.4	34.9	57.7	59.3	90.4	57.7	6.7	1.6	97.32

TABLE 4.3 DISCRETE INFORMATION SYSTEM

Sample	FC01	FC12	PC04	TD20	TD21	TD22	TD17	FC15	TD23	FC16	FC18	AN01
1	3	2	3	4	3	3	3	1	3	3	3	4
2	3	4	3	4	3	3	2	3	3	3	3	2
3	3	1	2	3	3	3	3	2	2	3	3	3
4	3	4	1	4	3	3	3	3	3	3	3	2
5	3	4	3	3	3	3	3	2	3	3	3	4
6	3	3	2	3	3	3	3	2	3	3	3	3
7	3	3	2	3	3	3	3	3	3	2	3	3
8	3	3	3	4	4	3	3	2	3	3	3	2
9	3	2	2	3	4	3	2	3	3	3	3	1
10	3	2	3	3	4	3	3	3	3	1	4	2
11	3	3	3	3	4	3	3	2	3	1	3	3
12	4	2	2	4	4	3	2	4	4	2	3	3
13	3	3	2	3	3	2	3	2	2	3	3	2
14	3	2	2	3	3	2	3	2	2	3	3	4
15	3	3	2	3	3	2	3	3	2	3	3	3
50	3	3	3	3	2	3	2	3	2	2	2	3

TABLE 4.4 DECISION TABLE AFTER COMBINATION AND DELETION

Sample	FC01	FC12	PC04	TD20	TD21	TD22	TD17	FC15	TD23	FC16	FC18	AN01
1	3	2	3	4	3	3	3	1	3	3	3	3
2	3	4	3	4	3	3	2	3	3	3	3	2
3	3	1	2	3	3	3	3	2	2	3	3	3
4	3	4	1	4	3	3	3	3	3	3	3	2
5	3	4	3	3	3	3	3	2	3	3	3	4
6	3	3	2	3	3	3	3	2	3	3	3	3
7	3	3	2	3	3	3	3	3	3	2	3	3
8	3	3	3	4	4	3	3	2	3	3	3	2
9	3	2	2	3	4	3	2	3	3	3	3	1
10	3	2	3	3	4	3	3	3	3	1	4	2
11	3	3	3	3	4	3	3	2	3	1	3	3
12	4	2	2	4	4	3	2	4	4	2	3	3
13	3	3	2	3	3	2	3	2	2	3	3	2
14	3	2	2	3	3	2	3	2	2	3	3	4
15	3	3	2	3	3	2	3	3	2	3	3	3
49	3	3	3	3	2	3	2	3	2	2	2	3

TABLE 4.5 DECISION TABLE WETHOUT FC02

Sample	FC01	PC04	TD112	TD21	TD22	TD17	FC15	TD23	FC16	FC18	AN01
1	3	3	4	3	3	3	1	3	3	3	3
2	3	3	4	3	3	2	3	3	3	3	2
3	3	2	3	3	3	3	2	2	3	3	3
4	3	1	4	3	3	3	3	3	3	3	2
5	3	3	3	3	3	3	2	3	3	3	4
6	3	2	3	3	3	3	2	3	3	3	3
7	3	2	3	3	3	3	3	3	2	3	3
8	3	3	4	4	3	3	2	3	3	3	2
9	3	2	3	4	3	2	3	3	3	3	1
10	3	3	3	4	3	3	3	3	1	4	2
11	3	3	3	4	3	3	2	3	1	3	3
12	4	2	4	4	3	2	4	4	2	3	3
13	3	2	3	3	2	3	2	2	3	3	2
14	3	2	3	3	2	3	2	2	3	3	4
15	3	2	3	3	2	3	3	2	3	3	3
49	3	3	3	2	3	2	3	2	2	2	3